PCIe® Re-timer in Data Centers Platforms

Alex Umansky

Chief Architect, PCIe Solutions

IT PL, Huawei

Disclaimer

<u>Presentation Disclaimer</u>: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

Agenda

- **Introduction**
- **System Integration Challenges**
- **Re-timer Affect on Latency and Performance**
- **Conclusions**

PCIe® Re-timers

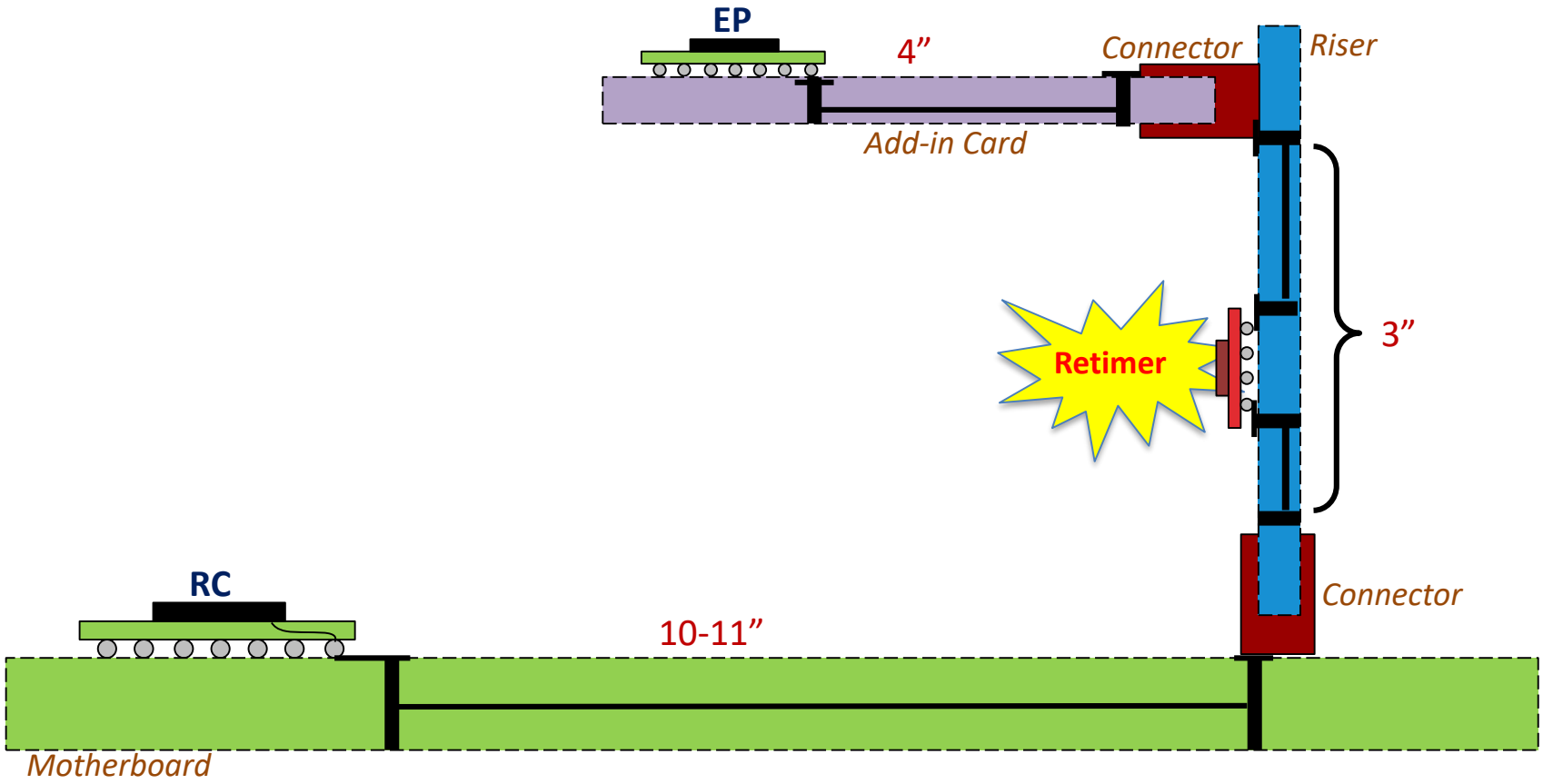
- **PCIe 4.0 is serious challenge from signal integrity point of view, especially for servers, storage and accelerator clusters platforms, where PCIe channel spending over relatively long traces, number of slots and connectors, cables and add-in cards**
- **Long and complicated channel exceeds -28db PCIe 4.0 spec defined limit**
- **Re-timers targeted to expend PCIe channel by implementing Ports with CDR and Equalization tuning capabilities**
- **While being transparent for Data/Transaction layers and above**

Range of Topologies

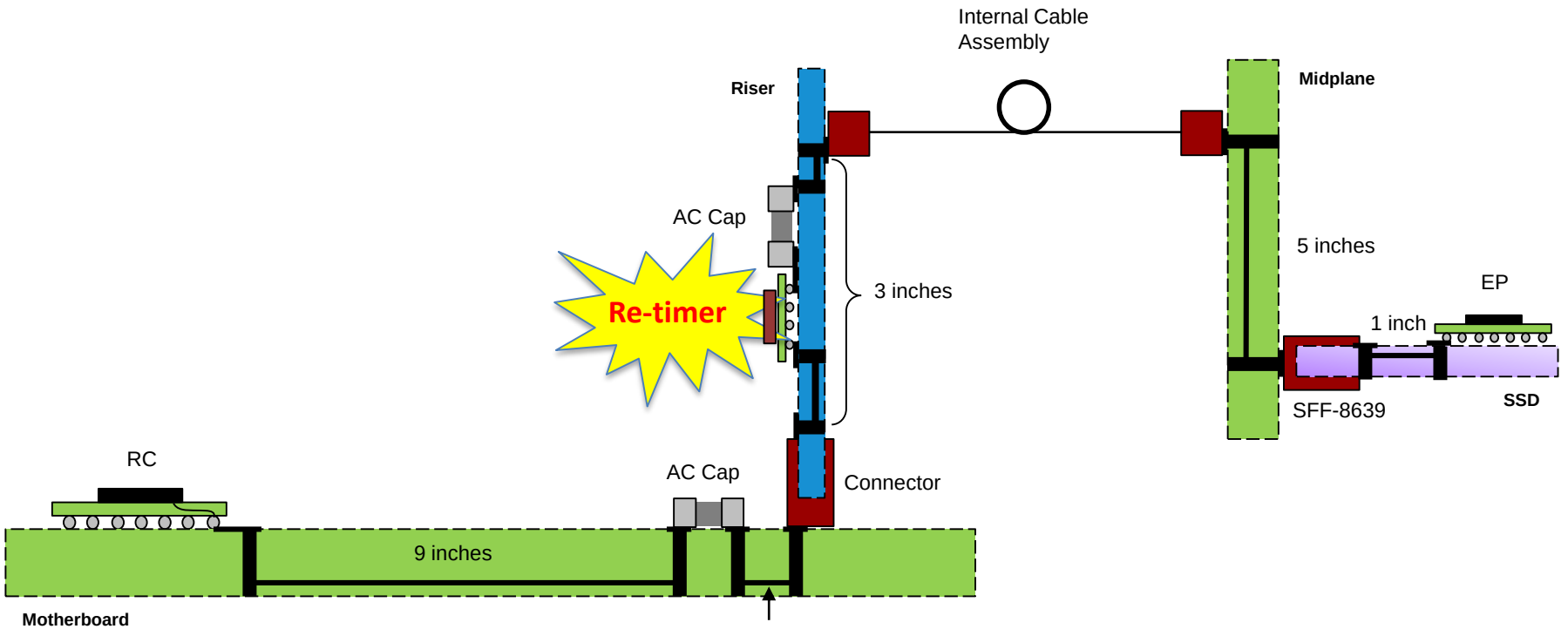
- **Server / Network FE / Storage Backplain / Accelerators Clusters / JBOF**

- **CPU ←RE-TIMER→ AIC**
- **CPU ←RE-TIMER→ Riser Card→ AIC**
- **CPU ←RE-TIMER→ Cable → Switch →AIC**
- **CPU → Switch ←RE-TIMER→ Cable →AIC**

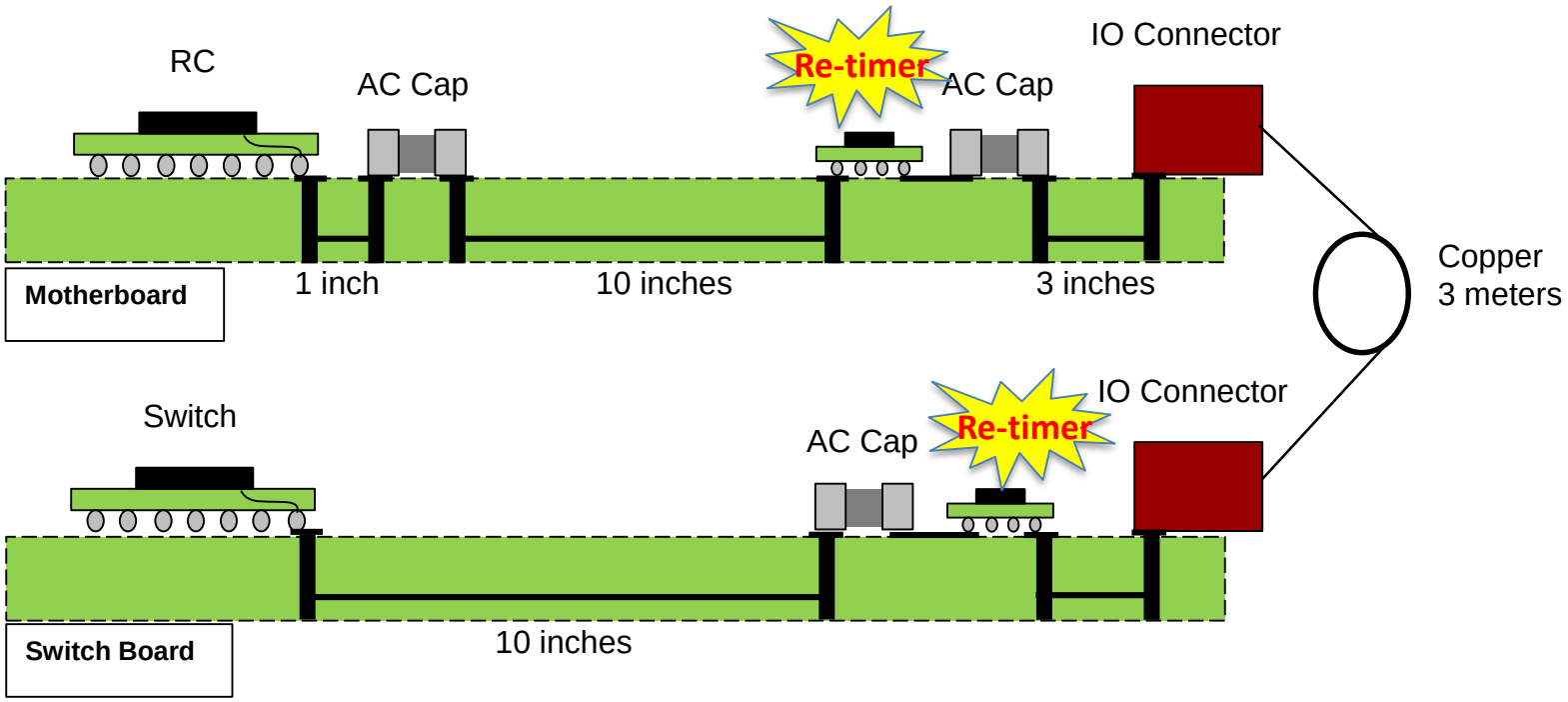
Re-timer Topology (1)



Re-timer Topology (2)



Re-timer Topology (3)



Re-timer Capabilities

- ✓ Physical Layer implemented
  - ✓ Phase 2/3 EQ procedure
  - ✓ EI and Inferred EI
  - ✓ RX impedance control on both side
- ✓ Totally transparent to DL and TL Layers
- ✓ TX: 3 stage FIR, full coefficient matrix
- ✓ Rx CDR
- ✓ RX DFE/CTLE
- ✓ Clock compensation by SKPOS add/remove
- ✓ Lane De-skew
- ✓ Lane numbering detection
- ✓ Lane polarity detection

- × No standard LTSSM
- × No CFG space
- × NO DL/TL /AL
- × No Error Handling . Errors are forwarded , including PHY /Symbol

- ○ x4 /x8 /x16 *V V V*
- ○ SRIS support on any pseudo port *V*
- ○ Receiver Lane margining *V*
- ○ Sideband registers rd/wr *V*
- ○ Slave Loopback *V*

Agenda

- Introduction
- **System Integration Challenges**
- Re-timer Affect on Latency and Performance
- Conclusions

System Challenges

LATENCY

- Spec defines 64+ ns re-timer latency
- Up to two re-timers on path
- 250+ns additional latency on IO—CPU RTT
- *Latency and performance effect on next slides*

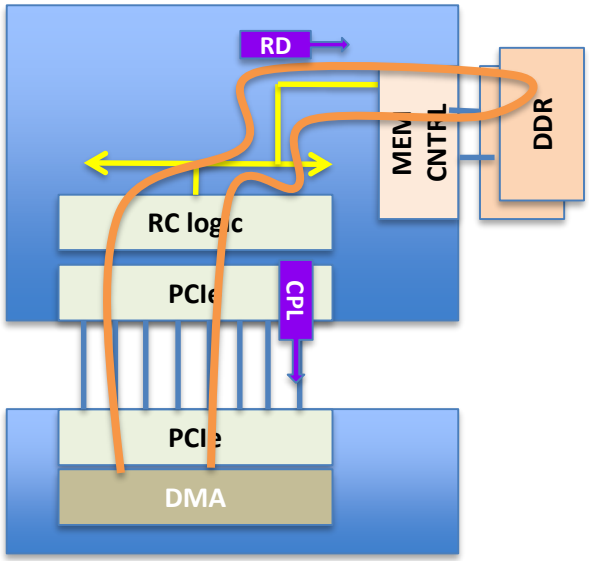
POWER

- Server CPU – 150W
- SSD – 30W
- x16 re-timer == 32 lanes → ~7W
- Additional 6-10% to platform power
- Heatsinks?

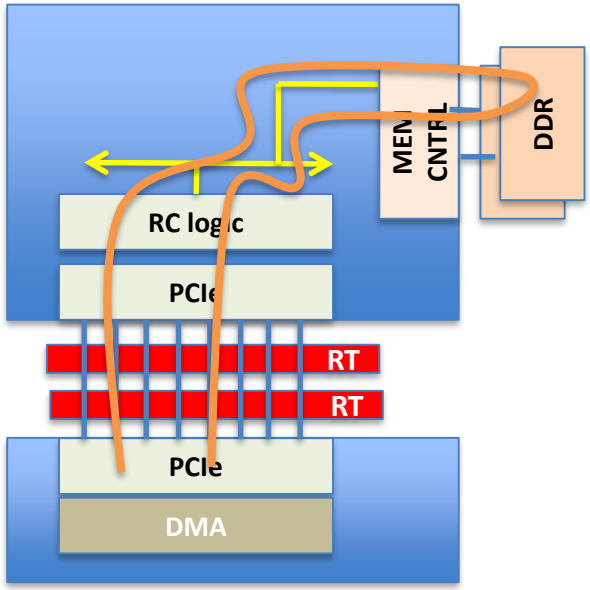
FOOTPRINT

- Additional component on board
- x2 link to SSD
- High density storage – M.2/BGA SSD

Latency

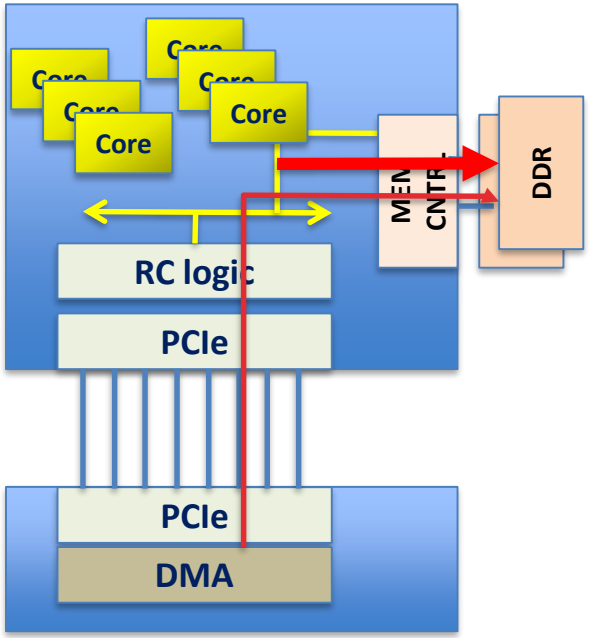


- ~250ns RTT



- ~250ns RTT + 256ns two re-timers
- +100% to RTT

Latency - Real Life Analyzes



RTT depend on dynamically change CPU load

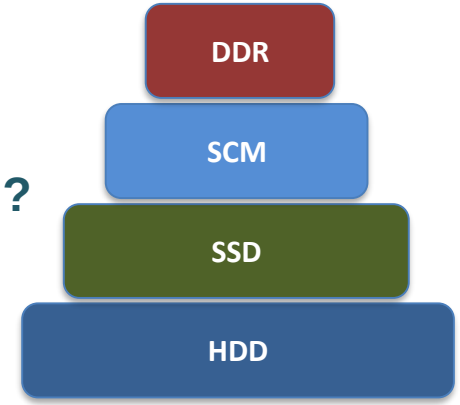
RTT may vary from 250ns to 3usec

Generally IO devices optimize for latency <1usec
Performance may be degraded for longer RTT

PCIe SSD – latency of 100us → 20us ?

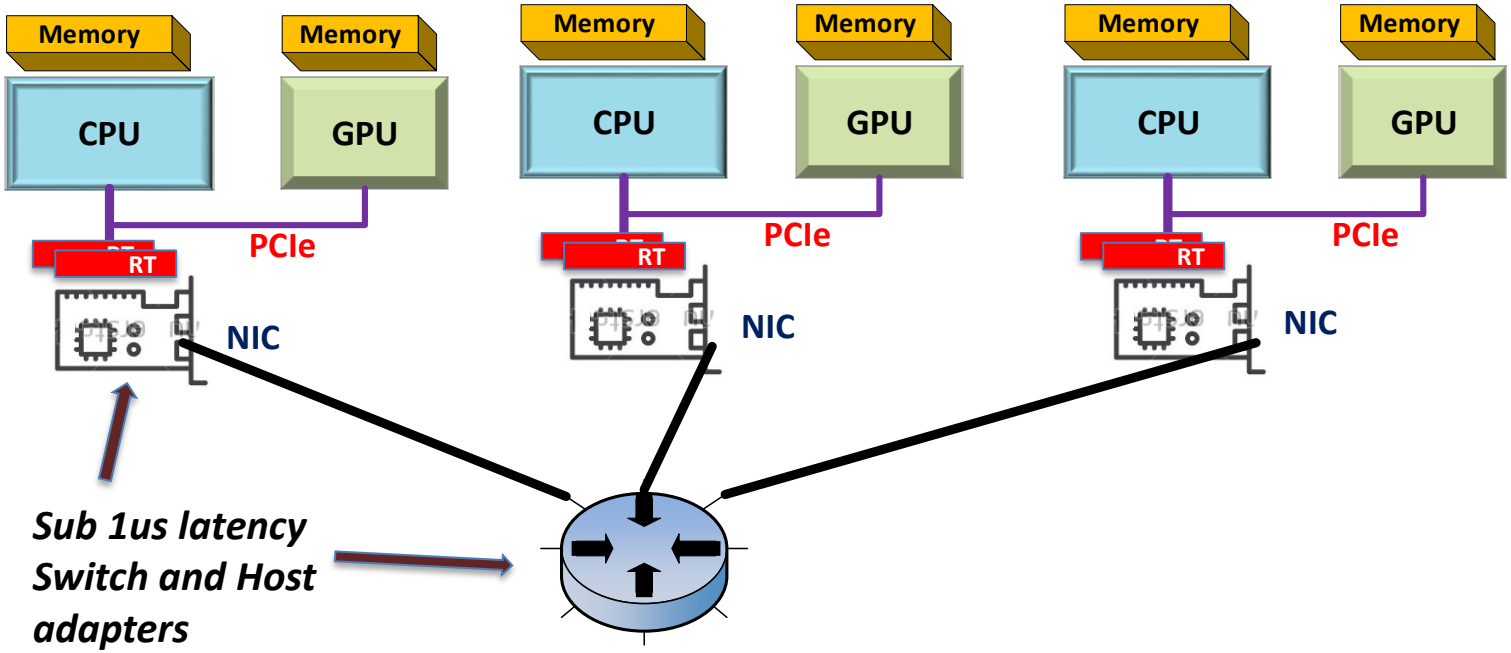
PCIe SCM ? Latency of 10us

PCIe FC HBA /NIC → HDDs (1-10ms)

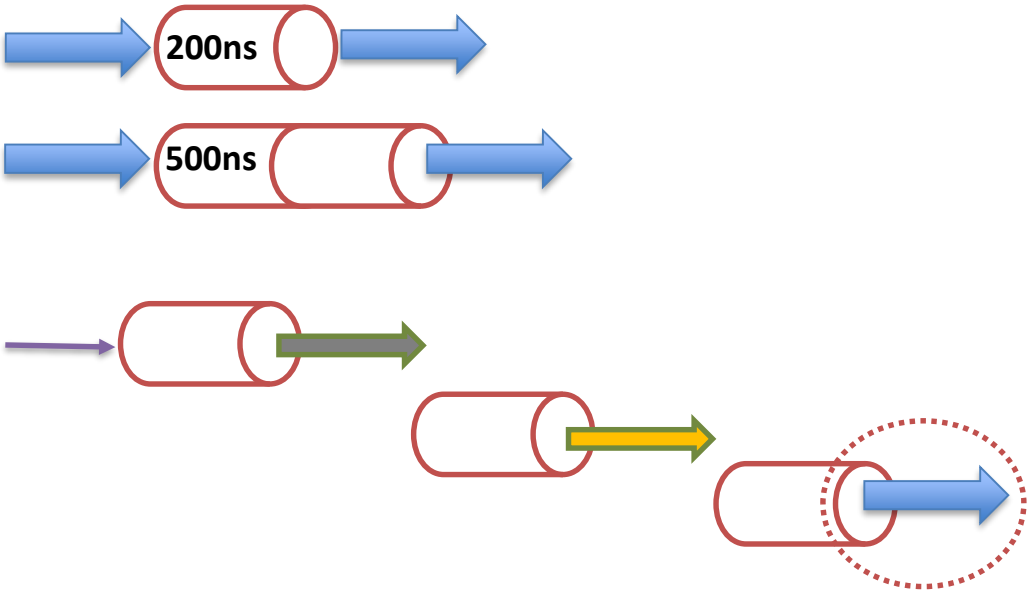


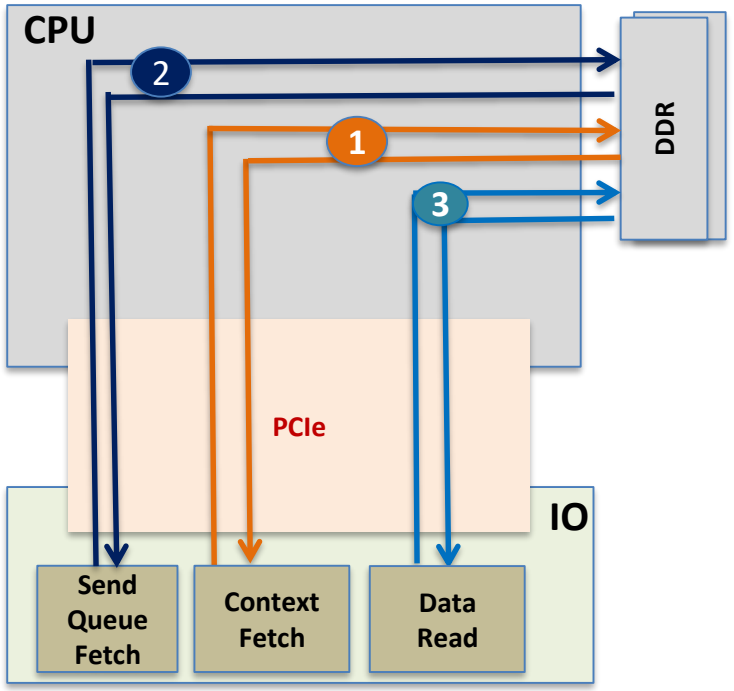
Latency in HPC Scenario

- Low latency networking (IB RDMA/RoCE) provides sub 1us latency
- Retimer latency will be increase this almost bare metal latency



Multi-stage IO Access





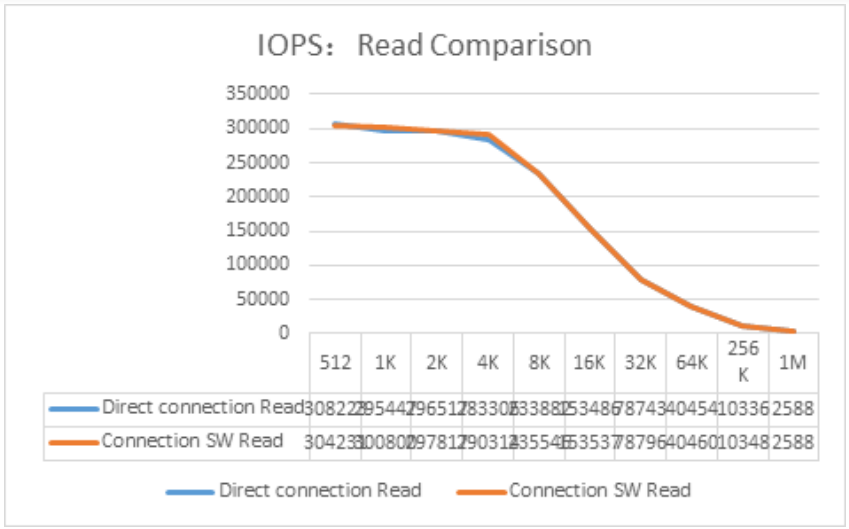
- Modern IO devices implementing Queues / Contexts /Tables in main memory
- Few accesses to main memory required for single chunk of data RX or TX operation
- Round trip latency may affect performance (IO/Packet rate)
- **IO try to prefetch Queues / Contexts / Data to mitigate RTT latency affect**

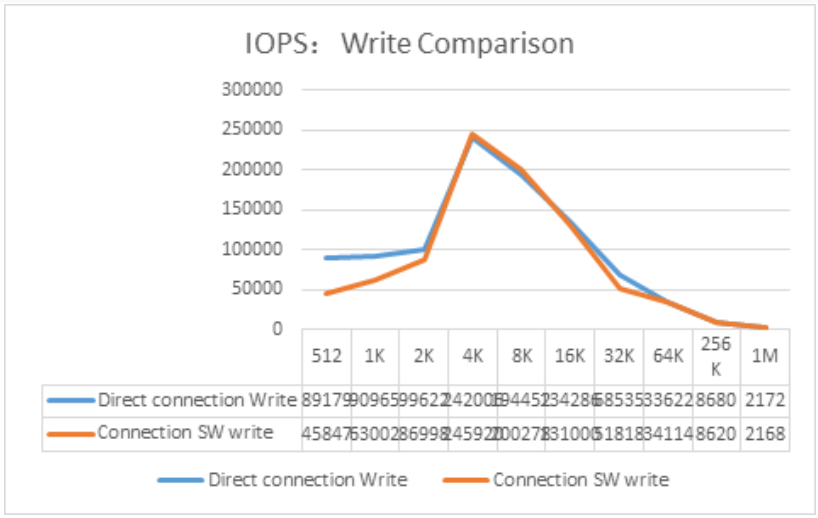
Performance vs. Latency (HCA)

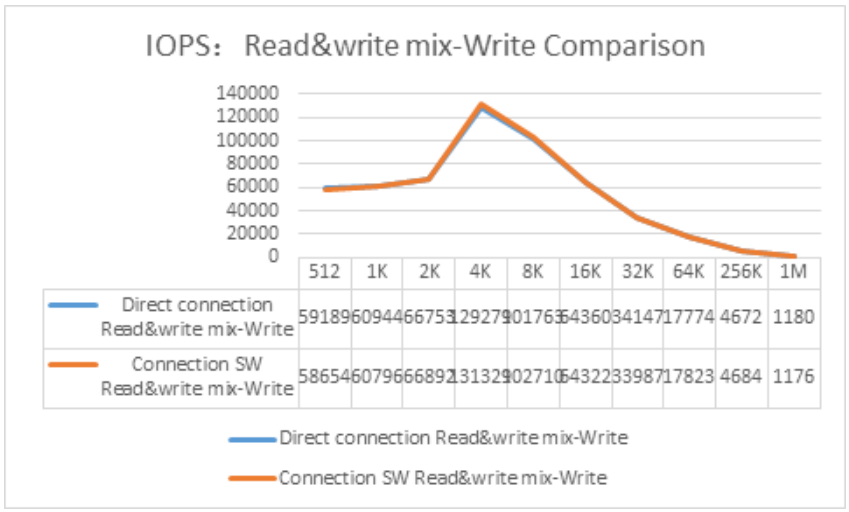
	64B	128B	256B	512B	768B	1K
Thru SW	9.01	9.01	9.01	8.1	6.6	5.18
Direct	9.01	9.017	9.017	8.96	7.2	5.797

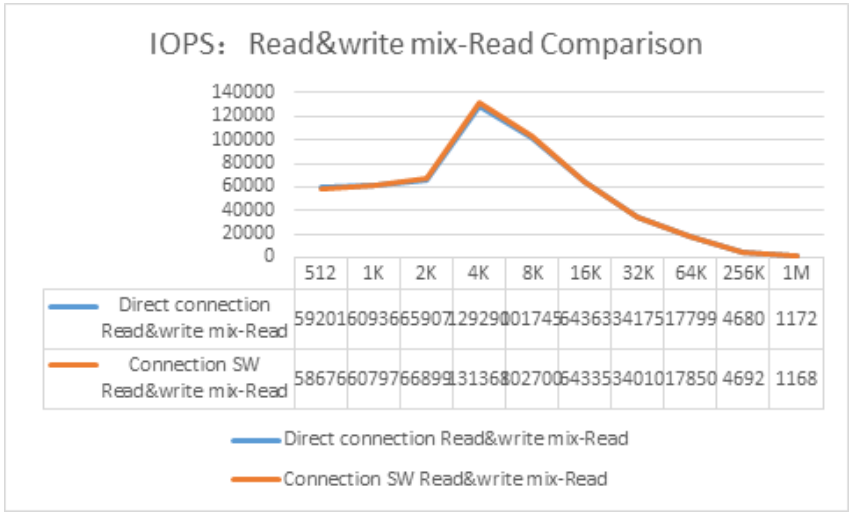
- **~170ns latency PCIe used to emulate re-timers latency**

Performance vs. Latency (SSD)

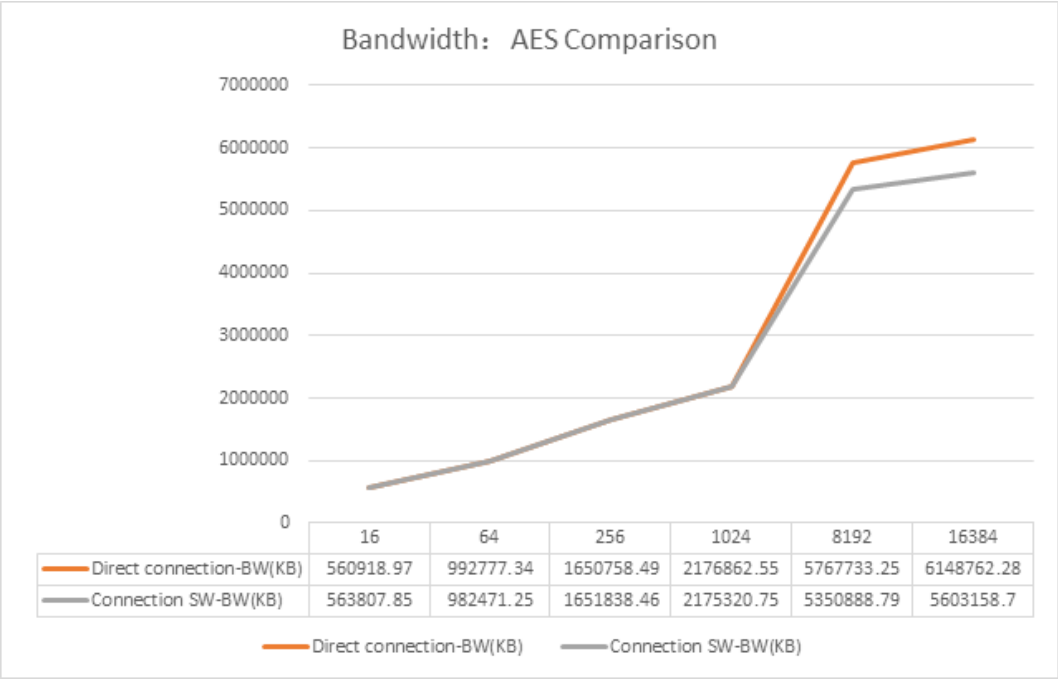








Performance vs. Latency (ACCLs)



rsa2048	TYPE	sign	verify
	Direct connection-OPS	15422.6	182011.8
	Connection SW-OPS	15433.4	168018.1

Conclusions

- **Standard re-timer may add 64+ns latency in one direction. This latency is critical for low latency applications.**
- **Ultra low (sub 10ns) latency solution required.**
- **Data and Transaction layers are ready for affectional latency (Scaled FC/10b Tag).**
- **Fine tuning of IO PCIe side parameters may be needed.**
- **IO buffers increase may be requested to mitigate additional latency effect.**

Thank you for attending the PCI-SIG Developers Conference 2018.

For more information, please go to www.pcisig.com

Don't forget to submit your feedback via the mobile app!

Download the **Crowd Compass** app and then search for **PCI-SIG Developers Conference** or entering the following URL into your mobile browser: https://crowd.cc/s/1rKy0
Enter event code: **DevCon2018**

Alternatively, access here: https://crowd.cc/pcisig2018

Note: Create an account within the app so Admin knows who to contact if selected as the prize winner.

Each session feedback is provided is equivalent to 1 raffle entry (up to 11 sessions).
General survey feedback = 1 raffle entry.